AMENDMENTS TO THE CLAIMS

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

Claim 1. (Currently amended) A computer-implemented method of determining cluster attractors for ~~use in clustering~~ a plurality of documents, each document comprising at least one term, each term comprising one or more words, the method comprising:

causing a computer to calculate, in respect of each term, a probability distribution ~~that is~~ indicative of

in the instance where a document comprises said term and one other term that co-occurs with said term in at least one of said documents, the frequency of occurrence of said one other term ~~in the instance where a document comprises said term and said one other term~~, and

in the instance where a document comprises said term and more than one other term that co-occurs with said term in at least one of said documents, the respective frequency of occurrence of each other term~~, that co-occurs with said term in at least one of said documents~~;

causing a computer to calculate, in respect of each term, the entropy of the respective probability distribution; and

causing the computer to select at least one of said probability distributions as a cluster attractor depending on the respective entropy value,

wherein the selected cluster attractor is a clustering focus for at least some of said documents.

Claim 2. (Original) A method as claimed in Claim 1, wherein each probability distribution comprises, in respect of each co-occurring term, an indicator that is indicative of the total number of instances of the respective co-occurring term in all of the documents in which the respective co-occurring term co-occurs with the term in respect of which the probability distribution is calculated.

Claim 3. (Previously presented) A method as claimed in Claim 1, wherein each probability distribution comprises, in respect of each co-occurring term, an indicator comprising a conditional probability of the occurrence of the respective co-occurring term in a document given the appearance in said document of the term in respect of which the probability distribution is calculated.

Claim 4. (Previously presented) A method as claimed in Claim 2, wherein each indicator is normalized with respect to the total number of terms in the document, or each document in which the term in respect of which the probability distribution is calculated appears.

Claim 5. (Original) A method as claimed in Claim 1, comprising assigning each term to one of a plurality of subsets of terms depending on the frequency of occurrence of the term; and selecting, as a cluster attractor, the respective probability distribution of one or more terms from each subset of

terms.

Claim 6. (Original) A method as claimed in Claim 5, wherein each term is assigned to a subset depending on the number documents of the corpus in which the respective term appears.

Claim 7. (Previously presented) A method as claimed in Claim 5, wherein an entropy threshold is assigned to each subset, the method comprising selecting, as a cluster attractor, the respective probability distribution of one or more terms from each subset having an entropy that satisfies the respective entropy threshold.

Claim 8. (Original) A method as claimed in Claim 7, comprising selecting, as a cluster attractor, the respective probability distribution of one or more terms from each subset having an entropy that is less than or equal to the respective entropy threshold.

Claim 9. (Previously presented) A method as claimed in Claim 5, wherein each subset is associated with a frequency range and wherein the frequency ranges for respective subsets are disjoint.

Claim 10. (Previously presented) A method as claimed in Claim 5, wherein each subset is associated with a frequency range, the size of each successive frequency range being equal to a constant multiplied by the size of the preceding frequency range in order of increasing frequency.

Claim 11. (Previously presented) A method as claimed in Claim 7, wherein the respective entropy threshold increases for successive subsets in order of increasing frequency.

Claim 12. (Original) A method as claimed in Claim 11, wherein the respective entropy threshold for successive subsets increases linearly.

Claim 13. (Canceled)

Claim 14. (Currently amended) An apparatus for determining cluster attractors for ~~use in clustering~~ a plurality of documents, each document comprising at least one term, each term comprising one or more words, the apparatus comprising:

means for calculating, in respect of each term, a probability distribution ~~that is~~ indicative of

in the instance where a document comprises said term and one other term that co-occurs with said term in at least one of said documents, the frequency of occurrence of said one other term ~~in the instance where a document comprises said term and said one other term~~, and

in the instance where a document comprises said term and more than one other term that co-occurs with said term in at least one of said documents, the respective frequency of occurrence of each other term;

means for calculating, in respect of each term, the entropy of the respective probability

distribution; and

means for selecting at least one of said probability distributions as a cluster attractor depending on the respective entropy value,

wherein the selected cluster attractor is a clustering focus for at least some of said documents.

Claim 15. (Currently amended) A computer-implemented method of clustering a plurality of documents, each document comprising at least one term, each term comprising one or more words, the method comprising:

causing a computer to calculate, in respect of each term, a probability distribution ~~that is~~ indicative of

in the instance where a document comprises said term and one other term that co-occurs with said term in at least one of said documents, the frequency of occurrence of said one other term ~~in the instance where a document comprises said term and said one other term~~, and

in the instance where a document comprises said term and more than one other term that co-occurs with said term in at least one of said documents, the respective frequency of occurrence of each other term~~, that co-occurs with said term in at least one of said documents~~;

causing a computer to calculate, in respect of each term, the entropy of the respective probability distribution;

causing the computer to select at least one of said probability distributions as a cluster attractor depending on the respective entropy value;

causing the computer to compare each document with each cluster attractor; and

causing the computer to assign each document to one or more cluster attractors depending on the similarity between the document and the cluster attractors,

wherein assigning each document to one or more cluster attractors creates a plurality of document clusters, each cluster comprising a respective plurality of documents.

Claim 16. (Original) A method as claimed in Claim 15, comprising: calculating, in respect of each document, a probability distribution indicative of the frequency of occurrence of each term in the document; comparing the respective probability distribution of each document with each probability distribution selected as a cluster attractor; and assigning each document to at least one cluster depending on the similarity between the compared probability distributions.

Claim 17. (Previously presented) A method as claimed in Claim 16, comprising organizing the documents within each cluster by: assigning a respective weight to each document, the value of the weight depending on the similarity between the probability distribution of the document and the probability distribution of the cluster attractor; comparing the respective probability distribution of each document in the cluster with the probability distribution of each other document in the cluster; assigning a respective weight to each pair of compared documents, the value of the weight depending on the similarity between the compared respective probability distributions of each

document of the pair; calculating a minimum spanning tree for the cluster based on the respective calculated weights.

Claim 18. (Canceled)

Claim 19. (Currently amended) An apparatus for clustering a plurality of documents, each document comprising at least one term, each term comprising one or more words, the apparatus comprising:

means for calculating, in respect of each term, a probability distribution ~~that is~~ indicative of

in the instance where a document comprises said term and one other term that co-occurs with said term in at least one of said documents, the frequency of occurrence of said one other term ~~in the instance where a document comprises said term and said one other term~~, and

in the instance where a document comprises said term and more than one other term that co-occurs with said term in at least one of said documents, the respective frequency of occurrence of each other term;

means for calculating, in respect of each term, the entropy of the respective probability distribution;

means for selecting at least one of said probability distributions as a cluster attractor depending on the respective entropy value;

means for comparing each document with each cluster attractor; and

means for assigning each document to one or more cluster attractors depending on the similarity between the document and the cluster attractors,

wherein assigning each document to one or more cluster attractors creates a plurality of document clusters, each cluster comprising a respective plurality of documents.

20. (New) A computer-implemented method of determining cluster attractors for use in clustering a plurality of documents, each document comprising at least one term, each term comprising one or more words, the method comprising:

causing a computer to calculate, in respect of each term, a probability distribution that is indicative of

in the instance where a document comprises said term and said one other term that co-occurs with said term in at least one of said documents, the frequency of occurrence of said one other term in the instance where a document comprises said term and said one other term, and

in the instance where a document comprises said term and more than one other term that co-occurs with said term in at least one of said documents, the respective frequency of occurrence of each other term, that co-occurs with said term in at least one of said documents;

causing a computer to calculate, in respect of each term, the entropy of the respective probability distribution; and

causing the computer to select at least one of said probability distributions as a cluster attractor depending on the respective entropy value,

wherein the selected cluster attractor is a clustering focus for at least some of said documents, and wherein said probability distribution is calculated as

$$p(y \mid z) = \frac{\sum_{x \in X(z)} tf(x, y)}{\sum_{x \in X(z), t \in Y} tf(x, t)}$$

where $tf(x, y)$ is a term frequency of a term y in a document x and $X(z)$ is a set of all documents of said plurality of documents that contain a term z and where t is a term index.

Claim 21. (New) A method as claimed in Claim 20, wherein said entropy is calculated as:

$$H(Y \mid z) = -\sum_{y} p(y \mid z) \log p(y \mid z)$$